

Estimating Probability Distributions using “Dirac” Kernels (via Rademacher-Walsh Polynomial Basis Functions) .

Hamse Y. Mussa and Avid M. Afzal*

EMIC Consultancy, 2 Stanley Avenue, Barking IG11 0LE, U.K.

** Centre for Molecular Sciences Informatics, Cambridge University, Lensfield Road, Cambridge CB2 1EW, UK*

mussax021@gmail.com; maa76@cam.ac.uk

Abstract

In many applications (in particular information systems, such as pattern-recognition, machine learning, cheminformatics, bioinformatics to name but a few) the assessment of uncertainty is essential – i.e., the estimation of the underlying probability distribution function. More often than not, the form of this function is unknown and it becomes necessary to non-parametrically construct/estimate it from a given sample.

One of the methods of choice to non-parametrically estimate the unknown probability distribution function for a given random variable (defined on binary space) has been the expansion of the estimation function in Rademacher-Walsh Polynomial basis functions.

In this paper we demonstrate that the expansion of the probability distribution function estimation in Rademacher-Walsh Polynomial basis functions is equivalent to the expansion of the function estimation in a set of “Dirac kernel” functions. The latter approach can ameliorate the computational bottleneck and notational awkwardness often associated with the Rademacher-Walsh Polynomial basis functions approach, in particular when the binary input space is large.

Keywords: *Binary spaces, Rademacher-Walsh, Dirac kernel function.*

1 Introduction

The assessment of uncertainty is important in quantitative science. This requires the estimation of the underlying probability distribution function explicitly or implicitly. However, the form of the function is usually unknown and it becomes necessary to non-parametrically construct/estimate the function from a given sample. When the random variable of interest is an L -dimensional binary “vector” (i.e., it resides in a binary space $\mathcal{B} = \{0, 1\}^L$), its L -dimensional probability distribution function $p(\mathbf{x})$ is often non-parametrically estimated

through 2^L Rademacher-Walsh Polynomial basis functions φ_i ¹ [1,2] as

$$p(\mathbf{x}) = \sum_{i=0}^{2^L-1} \alpha_i \varphi_i(\mathbf{x}) \quad (1)$$

where

$$\alpha_i = \frac{1}{2^L} \sum_{\mathbf{x} \in \mathcal{B}} p(\mathbf{x}) \varphi_i(\mathbf{x}) \quad (2)$$

The coefficients α_i can be estimated as [1]

$$\hat{\alpha}_i = \frac{1}{N} \sum_{j=1}^N \frac{1}{2^L} \varphi_i(\mathbf{x}_j) \quad (3)$$

where N refers to the number of available prototype patterns \mathbf{x}_j . Putting Eq. 3 into Eq. 1 yields [3]

$$\begin{aligned} \hat{p}(\mathbf{x}) &= \sum_{i=0}^{2^L-1} \frac{1}{N} \sum_{j=1}^N \frac{1}{2^L} \varphi_i(\mathbf{x}_j) \varphi_i(\mathbf{x}) \\ &= \frac{1}{N} \sum_{j=1}^N \sum_{i=0}^{2^L-1} \frac{\varphi_i(\mathbf{x}_j)}{\sqrt{2^L}} \frac{\varphi_i(\mathbf{x})}{\sqrt{2^L}} \\ &= \frac{1}{N} \sum_{j=1}^N K(\mathbf{x}_j, \mathbf{x}) \end{aligned} \quad (4)$$

where

$$K(\mathbf{x}_j, \mathbf{x}) = \sum_{i=0}^{2^L-1} \frac{\varphi_i(\mathbf{x}_j)}{\sqrt{2^L}} \frac{\varphi_i(\mathbf{x})}{\sqrt{2^L}} \quad (5)$$

For all practical purposes $L \ll \infty$; besides $\varphi_i(\mathbf{x}_j)$ (and $\varphi_i(\mathbf{x})$) can only take values 1, or -1 as illustrated in [1,2]. And according to [4,5], $K(\mathbf{x}_j, \mathbf{x})$ can be considered as a valid positive definite kernel function.

The estimation of $p(\mathbf{x})$ at \mathbf{x} can be instructively viewed as an average of how similar \mathbf{x} is to the given N prototype patterns \mathbf{x}_j , where $K(\mathbf{x}_j, \mathbf{x})$ is the similarity function [1,2,3,6]. If the available N prototype patterns constituting the

¹ According to Duda and Hart [1] this basis function set $\{\varphi_i(\mathbf{x})\}_{i=0}^{2^L-1}$ consists of a set of polynomials that can be generated by systematically forming the products of the distinct terms $2x_l - 1$ taken none at a time, one at a time, two at a time, three at a time, and so on, where $\mathbf{x} = (x_1, x_2, \dots, x_l, \dots, x_L)$. The resultant set is a complete set satisfying an orthogonality relation in their order – *i.e.*, $\varphi_i(\mathbf{x})$ and $\varphi_k(\mathbf{x})$ – with respect to the weighting function $w(\mathbf{x}) = 1$,

$$\sum_{\mathbf{x}} \varphi_i(\mathbf{x}) \varphi_k(\mathbf{x}) = \begin{cases} 2^L & i = k \\ 0 & i \neq k \end{cases}$$

where the summation is taken over all 2^L values of the binary “vectors”.

sample are distinct instances and $N = 2^L$, the estimated coefficients $\hat{\alpha}_i$ are exact [7]. However, exact or not, the expansion in Eq. 1 requires 2^L Rademacher-Walsh Polynomial basis functions, which can make the estimation notationally clumsy and computationally complicated whenever the value of L is large [1,8]. Thus, for Eq. 4 to have any practical use, knowledge of the closed form of the kernel function $K(\mathbf{x}_j, \mathbf{x})$ is essential. In the following section we demonstrate that the function $K(\mathbf{x}_j, \mathbf{x})$ in Eq. 5 is a “Dirac” kernel function [9]. Our concluding remarks are in the final section.

2 Main Idea

Here we present the nub of the paper: $K(\mathbf{x}_j, \mathbf{x})$ is a “Dirac” kernel function.

Theorem *If \mathbf{x} and $\mathbf{x}_j \in \mathcal{B}$, and $\varphi_i(\cdot)$ are Rademacher-Walsh Polynomial basis functions on \mathcal{B} , then*

$$K(\mathbf{x}_j, \mathbf{x}) = \sum_{i=0}^{2^L-1} \frac{\varphi_i(\mathbf{x}_j)}{\sqrt{2^L}} \frac{\varphi_i(\mathbf{x})}{\sqrt{2^L}} = \begin{cases} 1 & \mathbf{x}_j = \mathbf{x} \\ 0 & \mathbf{x}_j \neq \mathbf{x} \end{cases} \quad (6)$$

i.e., $K(\mathbf{x}_j, \mathbf{x})$ is a “Dirac kernel” function.

where $\mathbf{x}_j = \mathbf{x}$ means that $x_{j1} = x_1, x_{j2} = x_2, \dots, x_{jL} = x_L$, with x_{jl} and x_l referring to the binary-valued l^{th} elements of \mathbf{x}_j and \mathbf{x} , respectively.

As described in the Introduction, the set $\{\varphi_i(\mathbf{x})\}_{i=0}^{2^L-1}$ is obtained by systematically forming products of $(2x_l - 1)$ none at a time, one at a time, two at a time, three at a time, *etc.*, where $l = 1, 2, \dots, L$. By the same token the set $\{\varphi_i(\mathbf{x}_j)\varphi_i(\mathbf{x})\}_{i=0}^{2^L-1}$ is obtained by forming products of the distinct terms $(2x_{jl} - 1)(2x_l - 1)$ none at a time, one at a time, two at a time, three at a time, and so on.

Lemma 1 *Let a_1, a_2, \dots, a_L be L distinguishable real variables which can take the values of 1 and -1, and that their combinatorial compositions can be considered as products. The sum of their possible combinatorial compositions z_i , with $i = 0, 1, \dots, 2^L - 1$ is*

$$\sum_{i=0}^{2^L-1} z_i = \begin{cases} 2^L & \text{if } a_1, a_2, \dots, a_L = 1 \\ 0 & \text{if not} \end{cases} \quad (7)$$

Proof:

The possible combinations are the L variables chosen: no variable; 1 variable, a_i , at a time; 2 variables, $a_i a_j$, at a time; three variables, $a_i a_j a_k$, at a time; ...,; or L variables, $a_1 a_2 \dots a_L$, at a time.

If all the L variables are positive (Scenario 1), i.e., $a_k = +1$ (where $k = 1, 2, \dots, L$), then $z_0 = +1$ (when no variable is chosen); $z_1 = a_1 = +1, z_2 = a_2 = +1, \dots, z_L = a_L = +1; z_{L+1} = a_1 a_2 = +1, z_{L+2} = a_1 a_3 = +1, \dots, z_{L+\frac{L(L-1)}{2}} = a_{L-1} a_L = +1; \dots; \text{ and } z_{2^L-1} = a_1 a_2 \dots a_L = 1.$

Self-evidently the number of times that none of the variables is chosen is ${}^LC_0 = \binom{L}{0}$; the number of combinatorial terms containing one variable is ${}^LC_1 = \binom{L}{1}$; and the number combinatorial terms consisting of two, three, four, ..., and L variables are ${}^LC_\rho = \binom{L}{\rho}$, ρ being 2, 3, ..., and L , respectively. This means

$$\sum_{i=0}^{2^L-1} z_i = \sum_{\varrho=0}^L \binom{L}{\varrho} = 2^L \quad (8)$$

If all the L variables take the value of -1 (Scenario2), i.e., $a_k = -1$ where k is as defined before, then $z_0 = +1$; $z_1 = a_1 = (-1)^1$, $z_2 = a_2 = (-1)^1$, ..., $z_L = a_L = (-1)^1$; $z_{L+1} = a_1 a_2 = (-1)^2$, $z_{L+2} = a_1 a_3 = (-1)^2$, ..., $z_{L+\frac{L(L-1)}{2}} = a_{L-1} a_L = (-1)^2$; ..., and $z_{2^L-1} = a_1 a_2 \dots a_L = (-1)^L$. By the same token (as we reasoned above): ${}^LC_\varrho = (-1)^\varrho \binom{L}{\varrho}$ with $\varrho = 0, 1, 2, \dots, L$. Here

$$\sum_{i=0}^{2^L-1} z_i = \sum_{\varrho=0}^L (-1)^\varrho \binom{L}{\varrho}, \quad (9)$$

where obviously $\sum_{\varrho=0}^L (-1)^\varrho \binom{L}{\varrho} = 0$.

In the final scenario (Scenario3): For no specific reason, let us consider that m and k denote the number of variables that take the values -1 and 1, respectively, where $L = m + k$. In this scenario

$$\sum_{i=0}^{2^L-1} z_i = \sum_{\varrho=0}^{m+k} {}^{m+k}C_\varrho \quad (10)$$

It can readily be shown by induction that $\sum_{\varrho=0}^{m+k} {}^{m+k}C_\varrho = 0$ if one makes use of

these three identities:

I: ${}^nC_r = {}^{n-1}C_r + {}^{n-1}C_{r-1}$,

II: ${}^{n+r}C_{n+r} = {}^{n+r-1}C_{n+r-1}$, and

III: ${}^{n+j}C_0 = {}^{n+j-1}C_0$,

whereby r, j, n are non-negative integers and $r \leq n$ [10].

With $k = 1$, i.e., $\sum_{\varrho=0}^{m+k} {}^{m+k}C_\varrho$ becomes $\sum_{\varrho=0}^{m+1} {}^{m+1}C_\varrho$, which can be expressed as

$$\sum_{\varrho=0}^{m+1} {}^{m+1}C_\varrho = {}^{m+1}C_0 + \sum_{\varrho=1}^m {}^{m+1}C_\varrho + {}^{m+1}C_{m+1}$$

Making use of Identity I, the ${}^{m+1}C_\varrho$ on the RHS of the equation above becomes ${}^mC_\varrho + {}^mC_{\varrho-1}$, i.e., the equation can be rewritten as

$$\sum_{\varrho=0}^{m+1} {}^{m+1}C_{\varrho} = {}^{m+1}C_0 + \sum_{\varrho=1}^m {}^mC_{\varrho} + \sum_{\varrho=1}^m {}^mC_{\varrho-1} + {}^{m+1}C_{m+1},$$

which can be modified further by applying Identities III and II to the first and last terms on its RHS, respectively, resulting in

$$\sum_{\varrho=0}^{m+1} {}^{m+1}C_{\varrho} = {}^mC_0 + \sum_{\varrho=1}^m {}^mC_{\varrho} + \sum_{\varrho=1}^m {}^mC_{\varrho-1} + {}^mC_m = 2 \sum_{\varrho=0}^m {}^mC_{\varrho}$$

In *Scenario2* we have demonstrated that in the case that all the variables (denoted here by m) take the value of -1, ${}^mC_{\varrho} = (-1)^{\varrho} \binom{m}{\varrho}$. This means

$$\sum_{\varrho=0}^{m+1} {}^{m+1}C_{\varrho} = 2 \sum_{\varrho=0}^m (-1)^{\varrho} \binom{m}{\varrho} \quad (11)$$

In the case of $k=2$, $\sum_{\varrho=0}^{m+k} {}^{m+k}C_{\varrho}$ becomes $\sum_{\varrho=0}^{m+2} {}^{m+2}C_{\varrho}$, which can be expressed as

$$\sum_{\varrho=0}^{m+2} {}^{m+2}C_{\varrho} = {}^{m+2}C_0 + \sum_{\varrho=1}^{m+1} {}^{m+2}C_{\varrho} + {}^{m+2}C_{m+2}$$

Applying Identity I to ${}^{m+2}C_{\varrho}$ in the middle term on the RHS of the equation above, we obtain

$$\sum_{\varrho=0}^{m+2} {}^{m+2}C_{\varrho} = {}^{m+2}C_0 + \sum_{\varrho=1}^{m+1} {}^{m+1}C_{\varrho} + \sum_{\varrho=1}^{m+1} {}^{m+1}C_{\varrho-1} + {}^{m+2}C_{m+2}$$

By following the same line of reasoning as employed in the case of $k=1$ and applying Identities III and II to the first and last terms on the RHS of the equation above, respectively, gives

$$\sum_{\varrho=0}^{m+2} {}^{m+2}C_{\varrho} = 2 \sum_{\varrho=0}^{m+1} {}^{m+1}C_{\varrho}, \quad (12)$$

whereby $2 \sum_{\varrho=0}^{m+1} {}^{m+1}C_{\varrho}$ can be expressed as

$$2 \sum_{\varrho=0}^{m+1} {}^{m+1}C_{\varrho} = 2 \left[2 \sum_{\varrho=0}^m (-1)^{\varrho} \binom{m}{\varrho} \right] = 2^2 \sum_{\varrho=0}^m (-1)^{\varrho} \binom{m}{\varrho} \quad (13)$$

For $\sum_{\varrho=0}^{m+k} {}^{m+k}C_{\varrho}$, one just needs to repeat the process above k times, which gives

$$\sum_{\varrho=0}^{m+k} {}^{m+k}C_{\varrho} = 2^k \sum_{\varrho=0}^m (-1)^{\varrho} \binom{m}{\varrho}$$

In *Scenario2*, where $m = L$, it was shown that $\sum_{\varrho=0}^m (-1)^{\varrho} \binom{m}{\varrho} = 0$. Thus

$$\sum_{\varrho=0}^{m+k} {}^{m+k}C_{\varrho} = 2^k \sum_{\varrho=0}^m (-1)^{\varrho} \binom{m}{\varrho} = 0$$

I.e.

$$\sum_{i=0}^{2^L-1} z_i = \sum_{\varrho=0}^{m+k} {}^{m+k}C_{\varrho} = 0 \quad (14)$$

This finalizes the proof of **Lemma 1**.

3 Proof of Theorem 1

As described above, $\varphi_0(\mathbf{x}_j)\varphi_0(\mathbf{x}) = 1$ and the terms $\varphi_i(\mathbf{x}_j)\varphi_i(\mathbf{x})$ take the values +1 or -1, where $i = 1, 2, \dots, L$.

Now, if we consider $\varphi_1(\mathbf{x}_j)\varphi_1(\mathbf{x})$, $\varphi_2(\mathbf{x}_j)\varphi_2(\mathbf{x})$, ..., and $\varphi_L(\mathbf{x}_j)\varphi_L(\mathbf{x})$ as the real L variables in **Lemma 1**, then

$$\begin{aligned} z_0 &= \varphi_0(\mathbf{x}_j)\varphi_0(\mathbf{x}), \\ z_1 &= \varphi_1(\mathbf{x}_j)\varphi_1(\mathbf{x}), \\ &\cdot \\ &\cdot \\ &\cdot \\ z_{2^L-1} &= \varphi_{2^L-1}(\mathbf{x}_j)\varphi_{2^L-1}(\mathbf{x}) = [\varphi_1(\mathbf{x}_j)\varphi_1(\mathbf{x})][\varphi_2(\mathbf{x}_j)\varphi_2(\mathbf{x})] \dots [\varphi_L(\mathbf{x}_j)\varphi_L(\mathbf{x})]. \end{aligned}$$

Then by the virtue of **Lemma 1**,

$$\sum_{i=0}^{2^L-1} z_i = \sum_{i=0}^{2^L-1} \varphi_i(\mathbf{x}_j)\varphi_i(\mathbf{x}) = \begin{cases} 2^L & \text{if } \varphi_1(\mathbf{x}_j)\varphi_1(\mathbf{x}), \dots, \varphi_L(\mathbf{x}_j)\varphi_L(\mathbf{x}) = 1 \\ 0 & \text{if not} \end{cases} \quad (15)$$

Recall that the elements of the set $\{\varphi_i(\mathbf{x}_j)\varphi_i(\mathbf{x})\}_{i=1}^L$ take the value of 1 only if $\mathbf{x} = \mathbf{x}_j$. Multiplying on both side of Eq. 15 by $\frac{1}{\sqrt{2^L}} \frac{1}{\sqrt{2^L}}$ yields

$$\sum_{i=0}^{2^L-1} \frac{\varphi_i(\mathbf{x}_j)}{\sqrt{2^L}} \frac{\varphi_i(\mathbf{x})}{\sqrt{2^L}} = \begin{cases} 1 & \text{if } \mathbf{x}_j = \mathbf{x} \\ 0 & \text{if } \mathbf{x}_j \neq \mathbf{x} \end{cases} \quad (16)$$

which is Eq. 6 and this completes the proof of **Theorem 1**.

4 Conclusion

In this paper we have demonstrated that, on binary space \mathcal{B} , the expansion of the probability distribution estimation function in Rademacher-Walsh Polynomial basis functions is equivalent to the expansion of the estimation function in a set of Dirac kernel functions. The probability distribution estimation based on the Dirac kernel function scheme certainly alleviates both the computational bottle-necks and notational complexity associated with the Rademacher-Walsh Polynomial basis function approach, in particular when \mathcal{B} is large.

Acknowledgements It is a great pleasure to acknowledge Dr. J. B. O. Mitchell for reading the manuscript and his useful comments.

References

- [1] Duda, R. O. & Hart, P. E. J. (1973). *Pattern and Scene Analysis* (1st ed.), (Chapter 4). New York, US: John Wiley & Sons.
- [2] Hand, D. J. (1981). *Discrimination and Classification* (1st ed.), (pp. 106). Chichester, UK: John Wiley & Sons.
- [3] Meisel, W. S. (1972). *Computer-Oriented Approaches to Pattern Recognition* (1st ed.), (pp. 106). London, UK: Academic Press.
- [4] Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68, 337–404.
<http://www.ams.org/journals/tran/1950-068-03/S0002-9947-1950-0051437-7/>
- [5] Shawe-Taylor, J. & Cristianini N. (2004). *Kernel Methods for Pattern Analysis* (1st ed.), (pp. 60–66). Cambridge, UK: Cambridge University Press.
- [6] Parzen, E. (1962). On estimation of a probability density function and mode *Annals of Mathematical Statistics*, 33, 1065–1076.
<http://dx.doi.org/10.1214/aoms/1177704472>
- [7] Tou, J. R. & Gonzalez, R. C. (1974). *Pattern Recognition Principles* (1st ed.), (pp. 152–153). New York, US: Addison-Wesley
- [8] Hamse Y. Mussa, Jonathan D. Tyzack and Robert C. Glen (2013). Note on Rademacher-Walsh polynomials basis. *Journal of Mathematics of Research*, 5, 114–121.
- [9] Jacob, L. & Vert, J. (2008). Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, 24, 2149–2156.
<http://dx.doi.org/10.1093/bioinformatics/btn409>
- [10] Riley, K. F., Hobson, M. P. & Bence, S. J. (2007), *Mathematical Methods for Physics and Engineering*, (3rd ed.), (page 26). Cambridge, UK: Cambridge University Press.